(19)中华人民共和国国家知识产权局





(12)发明专利申请

(10)申请公布号 CN 109981485 A
(43)申请公布日 2019.07.05

(21)申请号 201910225762.4

(22)申请日 2019.03.25

(71)申请人 北京理工大学
地址 100081 北京市海淀区中关村南大街5号

(72)发明人 罗森林 王帅鹏 潘丽敏

(51)Int.Cl.
H04L 12/851(2013.01)
H04L 29/06(2006.01)
H04L 29/12(2006.01)

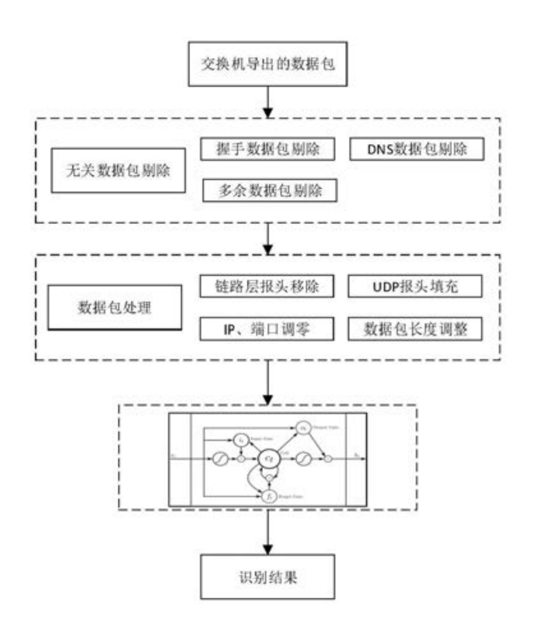
权利要求书1页 说明书3页 附图1页

(54)发明名称

基于长短期记忆网络的V2ray流量识别方法

(57)摘要

本发明涉及基于长短期记忆网络的V2ray流量识别方法，属于计算机网络安全领域。主要为了解决基于卷积神经网络的方法将数据转化为图片后训练出模型的可解释性较差，且未利用到加密流量在时间序列特征的问题。本发明首先从交换机获取V2ray流量和普通流量的数据链路层数据包并对数据包进行标注，其次去除不包含有用信息和冗余的数据包；然后将可能对模型训练造成影响的字节置零，对数据包的长度进行调整；最后使用这些预处理过的数据训练长短期记忆网络。该方法无需进行特征提取和选择，对V2ray流量的时间序列关系进行了学习，具有较好的识别效果。



CN 109981485 A

1.基于长短期记忆网络的V2ray流量识别方法,其特征在于所述方法包括如下步骤:

步骤1,从交换机设备中获得数据链路层数据包并标注为V2ray流量或其他流量;

步骤2,去除数据中不包含有用信息和冗余的数据包,去除TCP三次握手数据包,去除DNS域名解析数据包,保留每次通信的前16个数据包,并将这16个数据包作为数据集中的一条数据;

步骤3,对数据链路层数据包进行处理,去除数据链路层报头获得网络层数据包,对UDP报头进行填充使其长度与TCP报头保持一致,去除网络层数据报头中的表示IP地址和端口的信息,对数据包长度进行调整,使其保持一致;

步骤4,使用这些预处理过的数据训练长短期记忆网络。

2.根据权利要求1所述的基于长短期记忆网络的V2ray流量识别方法,其特征在于:步骤2中去除TCP三次握手数据包,去除DNS域名解析数据包,保留每次通信的前16个数据包。

3.根据权利要求1所述的基于长短期记忆网络的V2ray流量识别方法,其特征在于:步骤3将UDP报头补零扩充为20字节。

4.根据权利要求1所述的基于长短期记忆网络的V2ray流量识别方法,其特征在于:步骤3将TCP报头和UDP报头表示目的地址、目的端口、源地址、源端口的字节修改为0。

5.根据权利要求1所述的基于长短期记忆网络的V2ray流量识别方法,其特征在于:步骤3中通过补零和截断的方法将每个数据包的长度修改为1500字节。

基于长短期记忆网络的V2ray流量识别方法

技术领域

[0001]　本发明涉及基于长短期记忆网络的V2ray流量识别方法，属于计算机网络安全领域。

背景技术

[0002]　V2ray是一种新型的网络通信加密软件。其支持多种加密协议，并具有动态端口绑定、端口转发等功能，具有较高的灵活性、隐蔽性。目前对加密流量识别方法主要分为基于规则匹配的方法、基于机器学习的方法和基于深度学习的方法。

[0003]　1.基于规则匹配的方法

[0004]　基于规则匹配的方法通过对比数据库中的加密流量特征如端口信息、特定字节信息等识别加密通信软件。该方法步骤简单、判断过程极快，但端口转发、随机端口分配和流量伪装等技术的出现极大地降低了基于端口的识别方法的准确性。

[0005]　2.基于机器学习的方法

[0006]　基于机器学习的方法通过学习加密流量的统计特征达到对加密流量识别的目的，该方法具有较高的准确性，不依赖于一些可以被轻易改变的特征如端口号信息等。但基于机器学习的方法需要进行特征提取和特征选择，该过程时间成本和人工成本较高，且部分机器学习算法如K-NN分类器存在识别速率慢的问题。

[0007]　3.基于深度神经网络的方法

[0008]　基于深度学习的V2ray流量识别方法可以自动学习并提取加密流量中包含的特征信息，无需进行人工特征提取和选择，因而受到产业界的青睐，其中以卷积神经网络应用最为广泛。

[0009]　综上所述，近年来随着机器学习和深度学习技术的不断发展，越来越多的深度学习技术开始应用到计算机安全领域。现有的基于卷积神经网络的方法存在以下问题：(1)基于卷积神经网络的方法将数据转化为图片后训练卷积神经网络，模型的可解释性较差；(2)未利用到加密流量在时间序列上的特征。

发明内容

[0010]　本发明针对现有利用深度神经网络进行V2ray流量监测模型可解释性差、未利用V2ray流量在时间序列特征的问题，提出了基于长短期记忆网络的V2ray流量识别方法。

[0011]　本发明的技术方案是通过如下步骤实现的：

[0012]　步骤1，从交换机设备中获得数据链路层数据包并进行标注。

[0013]　步骤1.1，将这些数据包标记为V2ray流量或其他流量。

[0014]　步骤2，去除数据中不包含有用信息和冗余的数据包。

[0015]　步骤2.1，去除TCP三次握手数据包。

[0016]　步骤2.2，去除DNS域名解析数据包。

[0017]　步骤2.3，保留每次通信的前16个数据包，并将这16个数据包作为数据集中的一条

数据。

[0018] 步骤3,对数据链路层数据包进行处理。

[0019] 步骤3.1,去除数据链路层报头获得网络层数据包。

[0020] 步骤3.2,对UDP报头进行填充使其长度与TCP报头保持一致。

[0021] 步骤3.3,去除网络层数据报头中的表示IP地址和端口的信息。

[0022] 步骤3.4,对数据包长度进行调整,使其保持一致。

[0023] 步骤4,使用这些预处理过的数据训练长短期记忆网络。

[0024] 有益效果

[0025] 相比基于规则匹配的方法,本发明不依赖于端口特征和数据包内容特征,具有较低的误报率和漏报率。

[0026] 相比基于机器学习的方法,本发明无需进行特征提取和特征选择,降低了V2ray流量识别的复杂性和人工成本。

[0027] 相比基于卷积神经网络的方法,本发明可以对数据流时序关系进行记录和学习,提高了V2ray流量识别的准确率。

附图说明

[0028] 图1为本发明基于长短期记忆网络的V2ray流量识别方法原理图。

具体实施方式

[0029] 为了更好的说明本发明的目的和优点,下面对本发明方法的实施方式做进一步详细说明。

[0030] 1)所需数据均从交换机镜像端口获取。使用该方法获取到的数据包格式统一,与通信设备型号无关。且在部署到交换机设备上使用时无需对本方法进行额外的修改。获取到的数据需要标注为V2ray流量或其他流量。

[0031] 2)去除数据中不包含有用信息和冗余的数据包。TCP连接时为确保可靠性需要进行三次握手,三次握手过程中产生的SYN、ACK、FIN类型的TCP数据包不包含任何数据,无法为V2ray流量识别提供有用信息,这类数据包可以安全地剔除。DNS数据包负责进行域名解析,同样对流量监测没有帮助,应该剔除。

[0032] 3)V2ray服务端与客户端进行每次通信时需要预先交换密钥,因而每次通信较为靠前的数据包具有显著特征,其后所产生的数据包则为加密后的信息,内容较为随机。因而我们只保留每次通信的前16个数据包进行流量识别。

[0033] 4)从数据链路层获得的数据包报头为MAC地址信息,由设备不同而不同,需要去除。

[0034] 5)UDP报头长度为8字节,TCP报头长度为20字节,为了使数据包格式统一,将UDP报头补零扩充为20字节。

[0035] 6)TCP报头和UDP报头均包含目的地址、目的端口、源地址、源端口,在获取数据包的过程中,我们采用了数量有限的客户端和服务端,因而这些信息较为固定。为了使神经网络在训练过程中不学习到这些特征,应该将这些信息填充为0。

[0036] 7)深度神经网络需要长度固定的输入,由于互联网上的大部分数据包长度不超过

1500字节，因而我们通过补零和截断的方法将每个数据包的长度修改为1500字节。

[0037] 8)使用处理完成的数据训练长短期记忆网络，得到最终的模型。

[0038] 9)该模型按图1所示的原理图即可进行V2ray流量识别。

[0039] 以上所述的具体描述，对发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限定本发明的保护范围，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

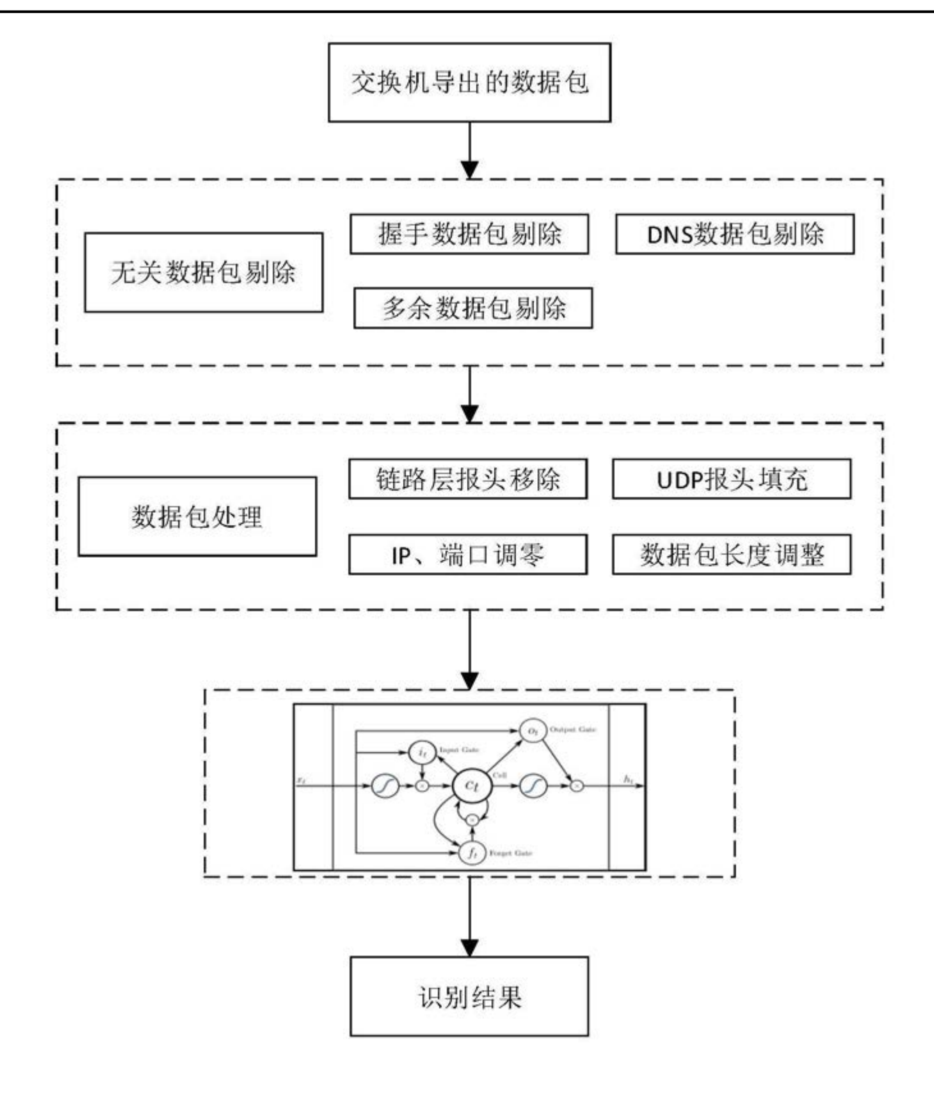

交换机导出的数据包
无关数据包剔除
握手数据包剔除
DNS数据包剔除
多余数据包剔除
数据包处理
链路层报头移除
UDP报头填充
IP、端口调零
数据包长度调整
x_t
i_t
Input Gate
o_t
Output Gate
c_t
Cell
h_t
f_t
Forget Gate
识别结果

图1